Review for the manuscript: "On the reliability of seasonal snow forecasts" by Vorobeva E., Orsolini Y., de Rosnay P., Day J., Senan R., Decremer D., Vitart F.

General comments:

In their manuscript, the authors provide a detailed analysis on the reliability of seasonal forecasts of snow from the SEAS5 model. The authors use reliability diagrams to analyse the forecast skill of low- (lower tercile) and high- (upper tercile) snow accumulation winters (in terms of SWE) using two independent verification datasets as reference for the forecasts skill verification process, namely the ERA5-Land and the ESA Snow-CCI version 4. The verification process against different observational dataset, which assigns each pre-defined land region to a specific Category of forecast reliability based on the definition of Weishemer and Palmer (2004), yields very different results. ERA5-Land generally yields higher reliability categories than ESA Snow-CCI, a discrepancy that the authors attribute in part to the shared dependence between the hindcasts and ERA5-Land on ERA5 forcing. In general, seasonal snow hindcasts are classified into "useful" categories (Categories 3–5) for both verification datasets, snow terciles and assessed regions.

The study compares the probability density functions for SWE anomalies across the two verification products. They find that while both datasets show similar general patterns, ERA5-Land typically displays greater variability and a higher frequency of extreme values. In contrast, ESA Snow-CCI anomalies tend to be more concentrated near the mean. Near the tercile thresholds, the agreement between the two datasets in classifying binary events (e.g., low- or high- snow accumulation occur) falls from an overall average of 70–80% to 50–70%. This suggests that the final reliability category assigned to a forecast is sensitive to the specific threshold definitions of the chosen reference data.

The overall study is rigorous and robust, the methodology is well explicated and therefore replicable, it addresses the very relevant and timely topic of reliability of seasonal forecasts of snow. However, I have some concerns regarding the overall structure of the manuscript, and the logic with which figures are presented. Hence, I suggest some major changes prior the manuscript could be accepted for publication.

A: We thank the reviewer for taking their time to review our manuscript. Please find a detailed reply to your comments below.

Specific comments:

In the Introduction, the discussion of previous literature on the verification of snow forecast products is limited. Expanding this section would help better contextualize the

present work within the broader literature and clarify its contribution relative to existing studies. For instance, in the Introduction the authors say: “Reliability of seasonal-mean near-surface temperature and precipitation forecasted by the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) 4 was examined in Weisheimer and Palmer (2014).” While I guess that the original reference was likely introduced to cite the study on which the methodology is based, there are a number of more recent studies that have evaluated the performance of the newer ECMWF seasonal forecast systems (SEAS5) through similar approaches. Some examples (also using the author’s methodology) are:

- Manzanas, , Torralba, V., Lledó, L., & Bretonnière, P. A. (2022). On the reliability of global seasonal forecasts: Sensitivity to ensemble size, hindcast length and region definition. Geophysical Research Letters, 49, e2021GL094662. https://doi.org/10.1029/2021GL094662
- Manzanas, R., Lucero, A., Weisheimer, A., & Gutiérrez, J. M. (2018). Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dynamics, 50*(3–4), 1161–1176. https://doi.org/10.1007/s00382-017-3668-z

A: We thank the anonymous reviewer for suggesting useful references. Those and other references have been added to introduction section to expand the discussion of previous literature.

In its current form, Figure 2 is introduced in relation to the subdivision into eight land regions (L169). It presents climatological mean winter SWE and standard deviation for three datasets, but its role in the manuscript somewhat unclear. The figure is only briefly discussed, with explicit reference mainly to panels (b), (d), and (f), while the climatological panels (a) and (e) are not described, and panel (c) is only mentioned in relation to the mountain mask (L176). A more complete description of all panels and a clearer link to the regional subdivision, including how it is adapted from Giorgi and Francisco (2000), would help to clarify its purpose in the manuscript.

A: We have explained the choice of land regions and clarified how those are related to the Giorgi and Francisco (2000). In addition, Figure 2 is now better described and cited.

The authors do not mention any detrending has been applied prior to the reliability assessment. If significant long-term trends are present in both the hindcasts and the verification datasets, these could affect he forecast skill and potentially masking the actual capability of the model to capture true interannual variability. While the domain-averaged time series in Figure 5 do not show obvious, pronounced trends, these large-scale averages may obscure significant regional trends occurring at the grid-point level. The authors may therefore consider assessing, or at least discussing, the sensitivity of

the reliability categories to detrending anomalies at each gridpoint, in order to ensure the robustness of the results.

A: Authors thank reviewer for this comment. In this manuscript, no detrending was applied to SWE anomalies prior to the reliability assessment, as was correctly noted. This choice was made because our evaluation aims to assess the forecast system in a real-world context, including any long-term systematic changes present in both hindcasts and observations. To our knowledge, none of the previously published reliability studies has mentioned detrending of anomalies. However, to address your concern that long-term trends could influence the results, we performed a sensitivity test in which SWE anomalies in both hindcast and observations were detrended prior to the reliability assessment. The results of this analysis are very similar to those originally shown, and all main conclusions remain the same, with little or no change in reliability category (see Figure below). For comparison with ERA5Land, only TIB region changed category from 4 to 3. However, we have already mentioned that this region is rather challenging. For comparison with ESA-CCI, even though some regions have changed their categories, the results are still showing marginally useful to marginally useful improved categories. This indicates that the reported forecast reliability is robust with respect to the presence or absence of long-term trends in the data. We have added this point to the revised manuscript.

Figure from the manuscript:

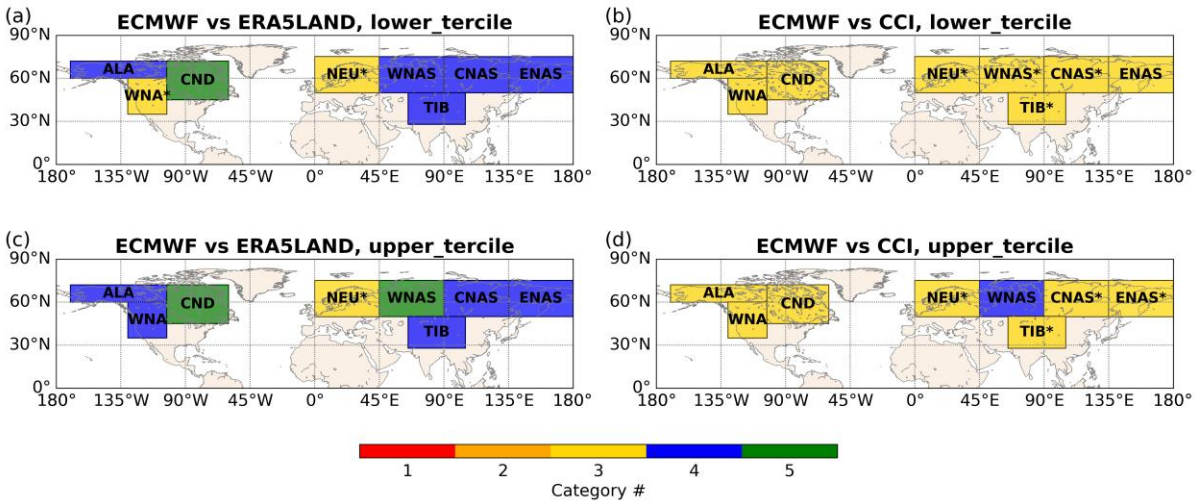
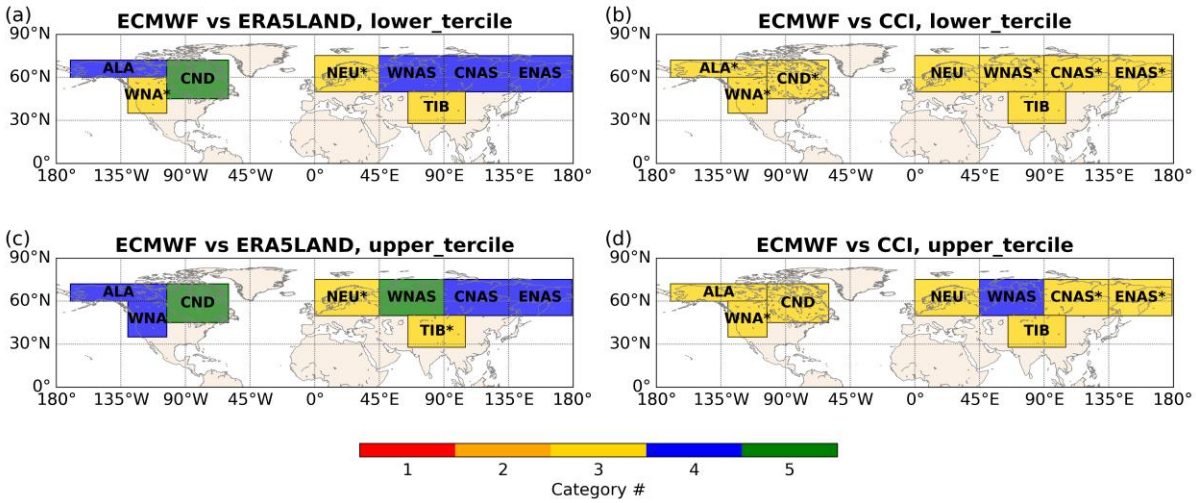


Figure for SWE anomalies detrended at each grid point:



I am a bit skeptical on the introduction of Fig. 4 and Fig. 5 only in the Discussion section. These figures could instead be presented and described in the Results section, with their implications for the reliability assessment discussed subsequently. Figure 4 is introduced as a means to better understand differences in reliability categories between the two verification datasets. However, as currently written, while the differences between the observational datasets anomalies are briefly described, their implications for the reliability assessment are not entirely clear to me. Although the authors mention the shared dependence between ECMWF forecasts and ERA5-Land on the ERA5 forcing, I think showing the ECMWF forecast distribution can provide additional insight into the behaviour of the forecasting system and help understanding possible bias in the forecasted anomalies distribution, better supporting the interpretation of the reliability differences found between verification datasets.

A: We thank reviewer for this point. Following the advice, we have expanded the results section which now contains Figures 3-8, including new Figures 6 and 8, and modified figures 4 and 5 (that were merged into Figure 7, and ECMWF anomalies were added). The figures are in the Results section and subsequently discussed.

In general, the structure of the manuscript needs to be improved, and the paper is lacking in discussion, with some paragraphs presenting results rather than providing interpretation. I have some further minor comments below that the authors should consider when revising their manuscript.

A: Thank you for providing valuable comments. The structure of the manuscript was improved, with an expansion of the Results section (as mentioned above), and a more focused Discussion section. Please find detailed replies below.

Technical corrections:

L8: The authors write “The results show good reliability of the ECMWF seasonal snow hindcasts for both low- and high-snow conditions.” However, throughout the manuscript, performance is consistently described in terms of “useful” or “marginally useful” categories. To be more specific, I would suggest adding a clarification such as “in that they consistently issue at least marginally useful SWE forecasts independently o the chosen benchmark”.

A: We have clarified the sentence, including a more precise listing of categories found for snow forecasts.

L10: “slightly higher reliability” can be made more precise like “1 or 2 categories better”.

A: The sentence has been clarified.

L30: I would smooth the sentence (e.g., “Improved forecasts of snow variables may also enhance the representation of large-scale atmospheric circulation in regions with strong snow–atmosphere coupling, such as East Asia”)

A: The suggestion is implemented in the text.

L43: “has”

A: Corrected

L45: The relevance of snow forecasts for hydrological applications is stated but not substantiated, could the authors include specific examples or relevant references?

A: This sentence has been removed as it repeated the idea expressed in the beginning of introduction section. One can find relevant references there.

L76: I think it would be useful to have some description or a reference explaining how UNC, REL and RES can be interpreted.

A: A reference to Hsu and Murphy (1986) is now added following Eq. 2.

L97: Figure 1 is introduced in the Methods section, to explain how to read a reliability diagram. However, in the manuscript, reliability diagrams themselves are never shown, but only appear in the Appenidx. It reproduces a standard schematic reliability diagram which is very similar to that presented in Weisheimer and Palmer (2014). I am not sure the figure adds value to the current manuscript, especially for a specialized audience. I suggest either explicitly stating that the figure is adapted from previous work and clarifying its purpose, or removing it and referring directly to the existing literature.

A: To satisfy the reviewer’s comment we have moved the reliability diagrams into the main body of the manuscript. We also believe that figure 1 does add value to the manuscript. It provides necessary introductory material (i.e., introducing key concepts and metrics) to readers who do not have background in forecast verification metrics as pointed out by the second reviewer. While it looks adapted from previous work, Figure 1

incorporates actual snow forecast data over western North Asia (WNAS) in the reliability diagram. We, therefore, leave figure 1 in the manuscript.

L192: This explanation seems to imply that the distribution of SWE anomalies is skewed, leading to a lower frequency of low-snow events. If so, I would indeed expect fewer forecasts with high probabilities for the lower tercile. Is my understanding what you mean? It would be useful to better clarify this point in the manuscript.

A: Thank you for the comment. We have now added a skewness coefficient information to Figure 7 and indeed, SWE distributions in all datasets are skewed. This indeed leads to fewer forecasts with high probabilities for the lower tercile experiment compared to the upper tercile experiment. However, we note that relatively small number of forecasts with high probabilities is a general feature of the forecast system and is observed for both lower and upper tercile experiments, as well as for the majority of regions (can be seen in individual diagrams). This indicates limited sharpness rather than an effect specific to the frequency of low-snow events. We have revised the text to clarify this point.

L201: While an offset of the best-fit reliability line from the climatological intersection appears to be present in both TIB and WNA regions, it is considerably more pronounced in TIB. In particular, in the TIB region the best-fit line lies systematically below the line of perfect reliability, indicating a tendency towards overforecasting. It would be helpful to further discuss this bias, as this provides useful insight into the nature of the forecast errors.

A: Thank you for this comment. We agree that the apparent offset of the best-fit reliability line in TIB and WNA regions suggests tendency towards overforecasting and requires further investigation. Upon closer examination of data in these regions, we found that this behavior was driven by grid points where observational time series exhibited no variability and were equal to zero SWE (no snow) throughout 30 years, while forecast timeseries did exhibit some variability. In such cases, the climatological threshold becomes effectively constant and equal to the observations themselves. This means that conditions (obs > threshold) or (obs < threshold) are never satisfied (upper and lower terciles in our study). As a result, these points contribute only to “no-event” outcomes across several forecast probability bins, which shifts the reliability line down. This is also the case for ERA5Land comparison, even though less pronounced due to fewer grid points with permanent snow-free conditions. We have added a clarification, a new figure and a discussion of this effect in the revised manuscript.

L213: I am not sure that repeatedly specifying the color of each category adds value. Once defined, this information could be omitted to improve readability and avoid unnecessary repetition.

A: Corrected accordingly.

L247: Are the agreement values reported in Table A1 computed by pooling all gridpoints and forecasted years together?

A: The overall agreement is obtained by pooling all grid points and forecasted years together, while agreement near the tercile is an additional condition related to the distance between the SWE anomaly and the tercile threshold. The agreement values are now shown in Figure 8 rather than in a table.

L254: Given the sensitivity of binary-event reliability to tercile thresholds and the reduced agreement between verification datasets near tercile boundaries, the authors may consider complementing the categorical reliability analysis with a continuous probabilistic verification metric (for example, the Continuous Ranked Probability Score).

A: Thank you for the comment. Following your advice, we have computed the CRPS to complement the reliability assessment. The results are shown in new Figure 8 and show that while CRPS is similar in both cases, Spearman's rank correlation coefficient between SWE anomaly in ERA5LAND and ESA Snow-CCI is low. It indicates that the two verification datasets do not classify the same years consistently as low and high snow accumulation winters. Since reliability is evaluated for binary tercile events, the forecast system is verified against different realizations of the binary event depending on the observational reference leading to differences in the reliability categories. We note that sensitivity of the binary-event reliability to probabilistic thresholds is natural. Two verification datasets of different origins are compared in this study, which show differences in trends, spatial and temporal variability (Mudryk et al., 2025). We therefore believe that differences in the reliability assessment are acceptable, as only useful categories (3-5) are nevertheless obtained. Also note that despite Manzanas et al (2022) concluded that the choice of verification dataset "does not greatly affect conclusions", their results show difference in 1-2 categories as well (see figures 2 and 3 in their supplement).

L258: Could the authors better specify what they mean by "snow dataset specifically suited for the forecast verification purposes"?

A: We clarified this point in the revised manuscript.

L267 and L273: The manuscript concludes that the forecasts show "overall good performance". However, many regions fall into Category 3 (marginally useful), especially when evaluated against ESA Snow-CCI. Given that Category 3 includes cases with limited skill and the results are strongly dependent on the verification dataset, the interpretation of "overall good performance" appears overstated.

A: Unlike in previous works for t2m and precipitation, reliability of seasonal SWE hindcasts shows only useful categories (3-5). We acknowledge that some regions exhibit limited skill and that reliability dependents on the verification dataset. We have

made corrections to the conclusions section and tried to be more specific than “overall good performance” by naming categories of reliability.